INST327: Database Design and Modelling

Section: 0106

Final Project Submission

10 May 2019

Group GM: Ryan Golub, Ali Namin, Sramika Thapa, Bryan Hwang, & Patrick Shields

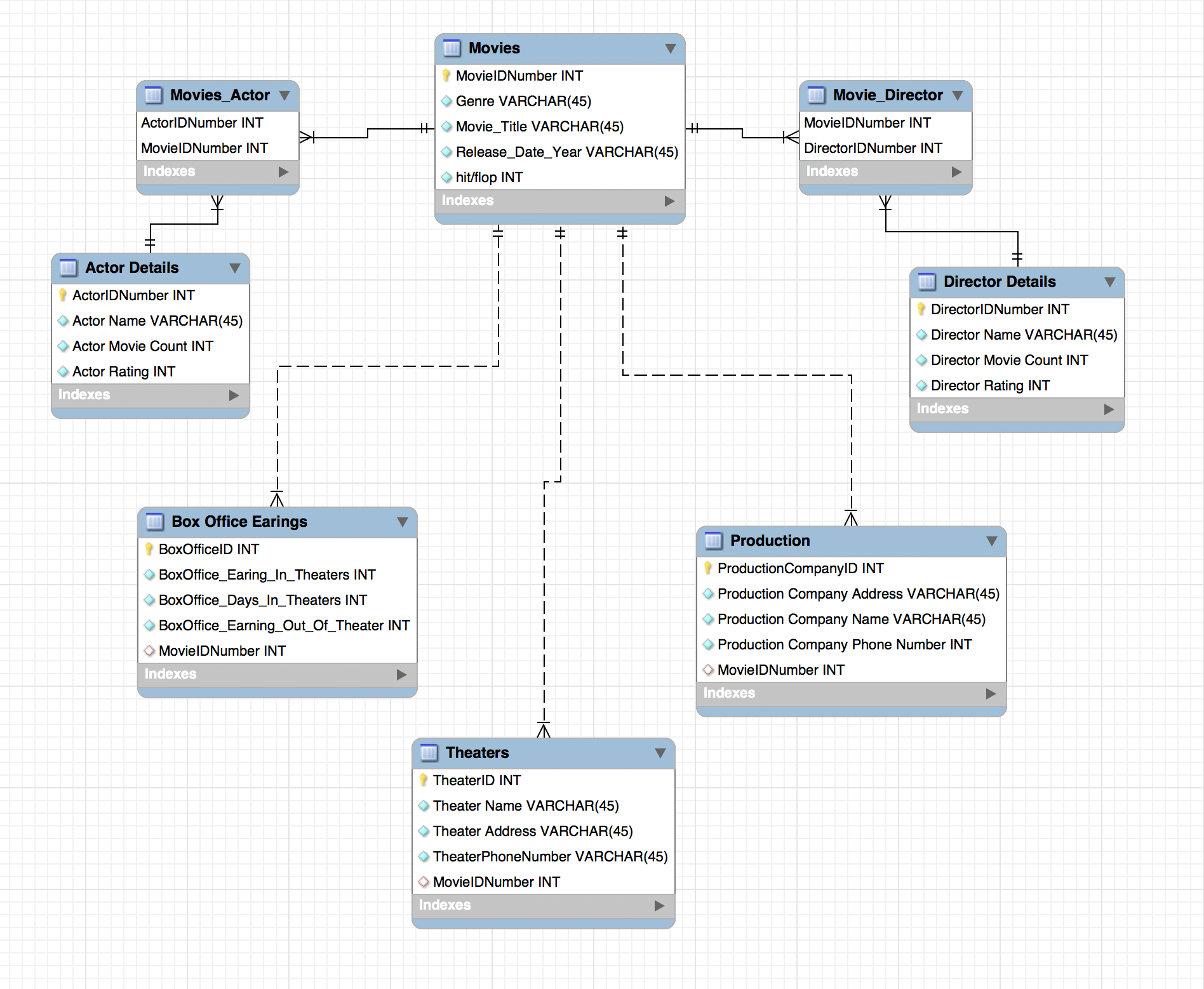
***Bollywood Movie Database***

Introduction

It’s got dancing, it’s got action, great stories, and interesting realism. It is Bollywood. When most people think Bollywood, they think of synchronized dance numbers and outrageous stunts. Bollywood actually has a plethora of genres that vary wildly. When attempting to find a movie to watch, the amount of movies Bollywood puts out each year can make it overwhelming. And although Wikipedia has a list of movies, it is only sorted by year, and not by any other category, which can make picking a movie feel like a game of roulette. In this way, many people can be deterred from attempting to watch these movies, and will miss out on the experience.

Our group aims to create a database of Bollywood movies that can be sorted by genre, director, main actors (male and female), and year released. We will cover movies from 2001 to 2005 so that we have a large dataset for those who are detail-oriented, but one that is still manageable for those who simply want quick information. In this way, we can fulfill the needs of people wishing to enter this multi-billion dollar fantasy world. We expect that a project like this can be extremely beneficial for people who want to watch past Bollywood movies that do not possess previous knowledge in that field. This can help to lower barrier to entry by helping people figure out which actors feature in which specific genres, and can also help those who look to research film development across different regions of filmmaking.

Database Description

O Logical Design (include a PNG of your ERD in this section)

O Physical Database



O Sample Data

Using Kaggle as a resource, we were able to gather a majority of our data from a

single csv file, which we later divided into smaller csv files to go into individual tables

as opposed to the ERD as a whole

O Views / Queries

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| View Name | Req. A | Req. B | Req. C | Req. D | Req. E |
| Query\_actorInMovies | X | X |  | X |  |
| Query\_actorMovieCount |  |  | X |  |  |
| Query\_directorAndProduction | X | X |  |  |  |
| Query\_topProduction\_companies | X | X |  |  | X |
| Query\_totalEarnings | X | X | X |  |  |

Changes from original design

Originally our group was thinking about using the entire database, but when it came time for our data import, we realized we had a lot of cutting down, rearranging, and organizing to do. Like a puzzle, we had to figure out which pieces of data belonged in which tables, which proved much more difficult than we had anticipated.

While the database downloaded from Kaggle had a lot of data, it still did not contain certain aspects that we still required for our database. So we filled in the CSV files to ensure no cells were empty with random information, such as Theater/Production address and Theater/Production phone number.

Lessons Learned

* Coordinating a group project can be difficult
* Normalization proved much more difficult than anticipated

This project challenged us to reevaluate our understanding of not only the subject matter, but also our abilities to work cooperatively. There were several occasions where the group realized that the amount of work demanded by the project scope was too much, and so revision of the goals achieved by the database had to be adjusted, sometimes with the advice of our designated TA. Additionally, the process of our ERD started out rather difficult for us when creating primary and foreign keys due to the problem of relationships between the tables. We did not realize until our TA meeting and a conversation with Professor Duffy that we had multiple joining tables because of the many to many relationships.

Potential Future Work

In our ERD, along with the data input, we were forced to use fake/made-up data in order to fulfill the requirement of 6 tables. Tables such as box office earnings, theaters, and production all were created using researched and made-up data. Currently the dataset we used were of movies that were in the early 2000s. If we were to continue this database and add to it, we would include more recent movies to help the user. Along with adding more recent movies, we would try to find a dataset that includes the tables of box office earnings, theaters, and production so we do not have to make up the data in the future.

References

Gulati, Suyash. “Bollywood 1920-2017.” *Bollywood 1920-2017*, Kaggle, 12 Jan. 2019, [www.kaggle.com/suyashgulati/bollywood-19202017/version/2](http://www.kaggle.com/suyashgulati/bollywood-19202017/version/2).

# Singh, Mitesh Kumar. *Bollywood Movie Dataset*, Kaggle, 10 June 2018, [www.kaggle.com/mitesh58/bollywood-movie-dataset#BollywoodMovieDetail.csv](http://www.kaggle.com/mitesh58/bollywood-movie-dataset#BollywoodMovieDetail.csv).

# 